

Web Data - CS 483

Whoosh Indexing Report

Super Clever Team Name

Aaron Goin
Giselle Gomez
Richie Zhang

Mar 10th 2017

Introduction

For Assignment 3 our team was to create a search index using Whoosh! over our previously collected datasets and to allow searching the index from the command-line. Our dataset was obtained from Wikipedia and comprised of all stubs contained within the Confections, Desserts, and Fruit categories. We wrote one script to scrape and parse the xml data into a SQL database. We then wrote another script to index the data with Whoosh! without stemming the attributes. To make up for the lack of stemming, we use query.Variations in our Query Parser which calculates word variations for every term the user enters and broadens our search net.

Database

Schema

Our SQL database consisted of three tables with identical schemas:

Category
Name: STRING
Image: STRING
Description: STRING
Source: STRING

Our Whoosh! indexed database combined the three tables into a single index with a simple schema:

Index
Name: TEXT
Category: TEXT
Image: TEXT
Description: TEXT
Source: TEXT

Both Image and Source are string URLs

Size

Our database contains 1077 tuples in total, with 403 Desserts, 372 Sweets, and 302 Fruits.

```

>>> from searcher import *
>>> main()

Opening existing index...

What would you like to search for? ("exit" to quit)
>? mango

Length of results: 11
Here are the first 10 results:

Name Category Image URL (may be shortened) Description (may be shortened) Source URL (may be
Angie (mango) Fruits https://upload.wikimedia.org/wikipedia/en/thumb/4/4c/Angie_ The Angie mango is a named mango cultivar that originated i https://en.wikipedi
Ivory (mango) Fruits https://upload.wikimedia.org/wikipedia/commons/thumb/4/40/M The Ivory, also called the jingu ivory, is a mango cultivar https://en.wikipedi
Anderson (mango) Fruits https://upload.wikimedia.org/wikipedia/commons/thumb/6/6f/M None https://en.wikipedi
Mango pomelo sago Desserts https://upload.wikimedia.org/wikipedia/commons/thumb/2/23/M Mango pomelo sago is a type of contemporary Hong Kong desse https://en.wikipedi
Dasherri Fruits https://upload.wikimedia.org/wikipedia/commons/thumb/2/2f/D Dasherri is a variety of mango grown in different parts of N https://en.wikipedi
Bennet Alphonso Fruits https://upload.wikimedia.org/wikipedia/commons/thumb/1/16/M The mango cultivar Bennet Alphonso is a derivative of the A https://en.wikipedi
Borong mangga Fruits https://upload.wikimedia.org/wikipedia/commons/thumb/8/8d/M Borong mangga is a Filipino food made by mixing sugar, salt https://en.wikipedi
Alampur Baneshan Fruits https://upload.wikimedia.org/wikipedia/commons/thumb/6/6c/M Alampur Benishan, often incorrectly spelt as Baneshan, is a https://en.wikipedi
Chok anan Fruits https://upload.wikimedia.org/wikipedia/commons/thumb/b/bf/U Chok Anan (Thai: ทุเรียน, pronounced [t͡ɕʰòːk ʔä.nān]) is https://en.wikipedi
Bilo-bilo Desserts https://upload.wikimedia.org/wikipedia/commons/thumb/f/fd/B Bilo-bilo is a Filipino dessert made of small gelatinous ba https://en.wikipedi

What would you like to search for? ("exit" to quit)
>?

```

Figure 0.1: Screenshot of a an example of a search. This searches the index for mango.

```

What would you like to search for? ("exit" to quit)
>? citrus OR cake OR chocolate

Length of results: 290
Here are the first 10 results:

Name Category Image URL (may be shortened) Description (may be shortened) Source URL (may be
Chocolate-covered cherry Sweets https://upload.wikimedia.org/wikipedia/commons/thumb/1/19/C Chocolate-covered cherries are a traditional dessert confec https://en.wikipedi
Chocolate bunny Sweets https://upload.wikimedia.org/wikipedia/commons/thumb/7/70/C A chocolate bunny or chocolate rabbit is a piece of chocola https://en.wikipedi
Flourless chocolate cake Desserts https://upload.wikimedia.org/wikipedia/commons/thumb/e/e6/F A flourless chocolate cake is a type of cake made from an a https://en.wikipedi
Snack cake Desserts https://upload.wikimedia.org/wikipedia/commons/thumb/6/67/H Snack cakes are a type of baked dessert confectionery made https://en.wikipedi
Amandine (dessert) Desserts https://upload.wikimedia.org/wikipedia/commons/thumb/e/e3/A Amandine is a Romanian chocolate layered cake filled with c https://en.wikipedi
Chocolate bullets Sweets https://upload.wikimedia.org/wikipedia/commons/thumb/7/70/C Chocolate bullets are a type of confectionery sold by confe https://en.wikipedi
Icebox cake Desserts https://upload.wikimedia.org/wikipedia/commons/thumb/c/cd/I An icebox cake (American) or chocolate ripple cake/log (Aus https://en.wikipedi
Chocolate coin Sweets https://upload.wikimedia.org/wikipedia/commons/thumb/8/81/C Chocolate coins, or chocolate money, are gold foil covered https://en.wikipedi
Wine cake Desserts https://upload.wikimedia.org/wikipedia/commons/thumb/4/43/W Wine cake, known in Spanish as torta envinada, is a cake ma https://en.wikipedi
Rum cake Desserts https://upload.wikimedia.org/wikipedia/commons/thumb/f/fc/R A rum cake is a type of dessert cake which contains rum. In https://en.wikipedi

What would you like to search for? ("exit" to quit)
>? citrus OR cake OR chocolate

```

Figure 0.2: Screenshot of a an example of a search. This searches the index for citrus OR cake OR chocolate.